HWG | HARRIS, WILTSHIRE & GRANNIS LLP

November 28, 2017

Ex Parte

Ms. Marlene H. Dortch
Secretary
Federal Communications Commission
445 12th Street, SW
Washington, DC 20554

Re: *Telecommunications Relay Services and Speech-to-Speech Services for Individuals with Hearing and Speech Disabilities*, CG Docket No. 03-123; *Misuse of Internet Protocol (IP) Captioned Telephone Service*, CG Docket No. 13-24

Dear Ms. Dortch:

As the Commission considers the potential for fully automated speech recognition to substitute for human Communications Assistants in the provision of IP CTS, it is important for the Commission to receive as much information as possible as to the evolution of automated speech recognition technology, and the path for future development. In that regard, CaptionCall, LLC, and its affiliate Sorenson Communications, LLC (together "CaptionCall"), provide the attached two papers.

The first, an October 11, 2017 post entitled "Speech Recognition is Not Solved," by Awni Hannun, provides an overview of the differences between automated speech recognition today, even with improved performance using the Switchboard database, and human-level speech recognition.[1]

The second article, entitled "Automated Speech Recognition for Captioned Telephone Conversations" and authored by Jeff Adams CEO, Cobalt Speech and Language, Inc., Kenneth Basye PhD, Alok Parlikar PhD, Andrew Fletcher PhD, and Jangwon Kim PhD, more specifically

[1] This blog is also located at: https://awni.github.io/speech-recognition/ (last viewed November 26, 2017).

addresses the challenges of developing fully automated automatic speech recognition for use in IP CTS.[2]

Sincerely,

John T. Nakahata
Counsel to CaptionCall, LLC and Sorenson Communications, LLC

cc: Nicholas Degani
Claude Aiken
Zenji Nakazawa
Amy Bender
Travis Litman
Nathan Eagan
Karen Peltz Strauss
Robert Aldrich
Eliot Greenwald
Michael Scott
Susan Bahr

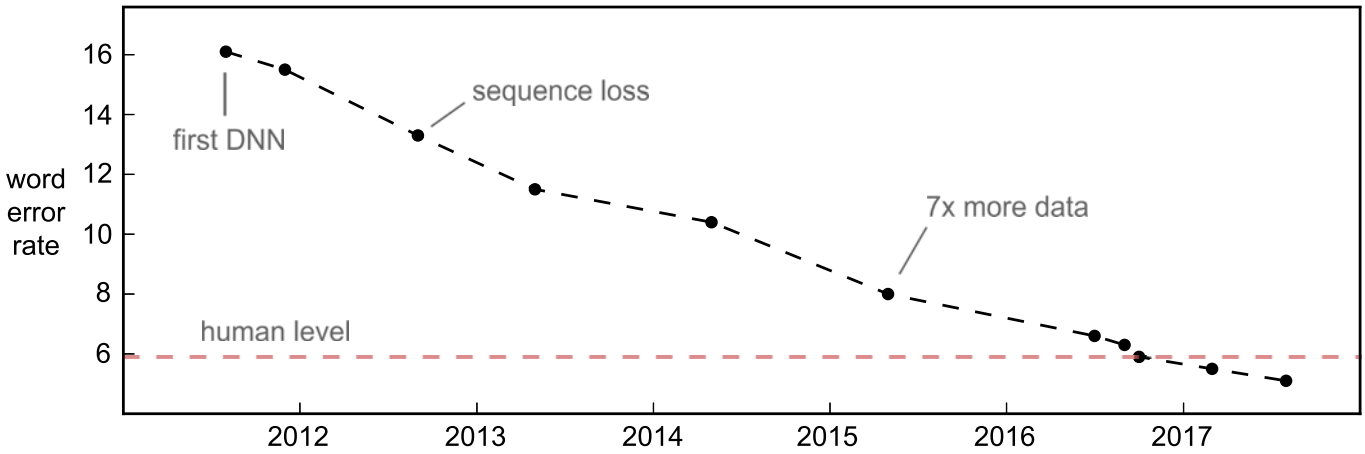
[2] This article is also available at: http://commons.clarku.edu/facultyworks/26 (last viewed November 26, 2017). While it was sponsored by CaptionCall, Inc., the thoughts and opinions expressed in the paper represent the independent views of the authors, who are leading academics and practitioners in the field.

Attachment 1

Speech Recognition Is Not Solved

Posted on October 11, 2017

Ever since Deep Learning hit the scene in speech recognition, word error rates have fallen dramatically. But despite articles you may have read, we still don't have human-level speech recognition. Speech recognizers have many failure modes. Acknowledging these and taking steps towards solving them is critical to progress. It's the only way to go from ASR which works for *some people, most of the time* to ASR which works for *all people, all of the time*.



Improvements in word error rate over time on the Switchboard conversational speech recognition benchmark. The test set was collected in 2000. It consists of 40 phone conversations between two random native English speakers.

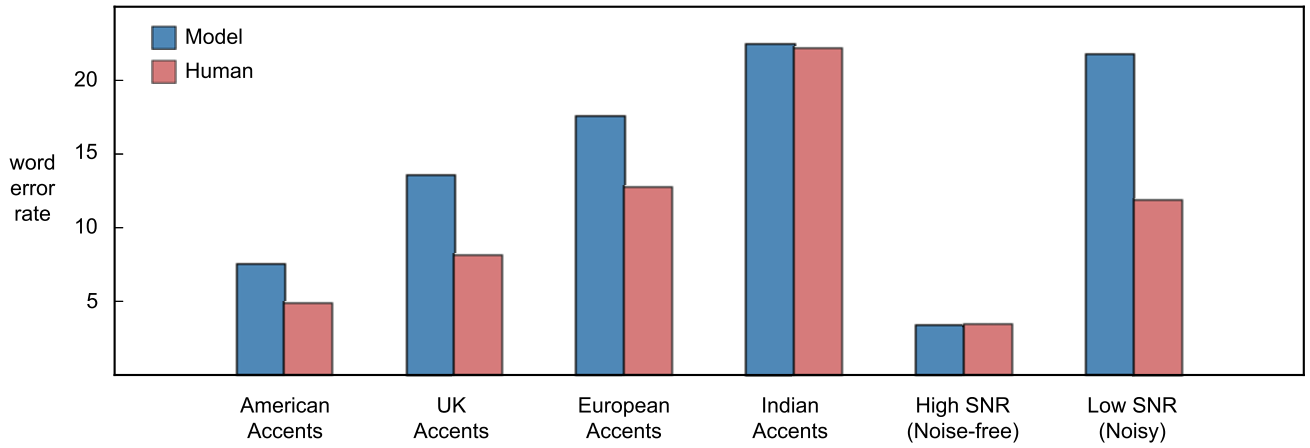
Saying we've achieved human-level in conversational speech recognition based just on Switchboard results is like saying an autonomous car drives as well as a human after testing it in one town on a sunny day without traffic. The recent improvements on conversational speech are astounding. But, the claims about human-level performance are too broad. Below are a few of the areas that still need improvement.

Accents and Noise

One of the most visible deficiencies in speech recognition is dealing with accents[1] and background noise. The straightforward reason is that most of the training data consists of American accented English with high signal-to-noise ratios. For example, the Switchboard conversational training and test sets only have native English speakers (mostly American) with little background noise.

But, more training data likely won't solve this problem on its own. There are a lot of languages many of which have a lot of dialects and accents. It's not feasible to collect enough annotated data for all

cases. Building a high quality speech recognizer just for American accented English needs upwards of 5 thousand hours of transcribed audio.



Comparison of human transcribers to Baidu's Deep Speech 2 model on various types of speech.[2] Notice the humans are worse at transcribing the non-American accents. This is probably due to an American bias in the transcribers. I would expect transcribers native to a given region to have much lower error rates for that region's accents.

With background noise, it's not uncommon for the SNR in a moving car to be as low as -5dB. People don't have much trouble understanding one another in these environments. Speech recognizers, on the other hand, degrade more rapidly with noise. In the figure above we see the gap between the human and the model error rates increase dramatically from the low SNR to the high SNR audio.

Semantic Errors

Often the word error rate is not the actual objective in a speech recognition system. What we care about is the *semantic error rate*. That's the fraction of utterances in which we misinterpret the meaning.

An example of a semantic error is if someone said "let's meet up Tuesday" but the speech recognizer predicted "let's meet up today". We can also have word errors without semantic errors. If the speech recognizer dropped the "up" and predicted "let's meet Tuesday" the semantics of the utterance are unchanged.

We have to be careful when using the word error rate as a proxy. Let me give a worst-case example to show why. A WER of 5% roughly corresponds to 1 missed word for every 20. If each sentence has 20 words (about average for English), the sentence error rate could be as high as 100%. Hopefully the mistaken words don't change the semantic meaning of the sentences. Otherwise the recognizer could misinterpret every sentence even with a 5% WER.

When comparing models to humans, it's important to check the nature of the mistakes and not just look at the WER as a conclusive number. In my own experience, human transcribers tend to make fewer and less drastic semantic errors than speech recognizers.

Researchers at Microsoft recently compared mistakes made by humans and their human-level speech recognizer.[3] One discrepancy they found was that the model confuses "uh" with "uh huh" much more frequently than humans. The two terms have very different semantics: "uh" is just filler

whereas “uh huh” is a *backchannel* acknowledgement. The model and humans also made a lot of the same types of mistakes.

Single-channel, Multi-speaker

The Switchboard conversational task is also easier because each speaker is recorded with a separate microphone. There’s no overlap of multiple speakers in the same audio stream. Humans on the other hand can understand multiple speakers sometimes talking at the same time.

A good conversational speech recognizer must be able to segment the audio based on who is speaking (*diarisation*). It should also be able to make sense of audio with overlapping speakers (*source separation*). This should be doable without needing a microphone close to the mouth of each speaker, so that conversational speech can work well in arbitrary locations.

Domain Variation

Accents and background noise are just two factors a speech recognizer needs to be robust to. Here are a few more:

- Reverberation from varying the acoustic environment.
- Artefacts from the hardware.
- The codec used for the audio and compression artefacts.
- The sample rate.
- The age of the speaker.

Most people wouldn’t even notice the difference between an `mp3` and a plain `wav` file. Before we claim human-level performance, speech recognizers need to be robust to these sources of variability as well.

Context

You’ll notice the human-level error rate on benchmarks like Switchboard is actually quite high. If you were conversing with a friend and they misinterpreted 1 of every 20 words, you’d have a tough time communicating.

One reason for this is that the evaluation is done *context-free*. In real life we use many other cues to help us understand what someone is saying. Some examples of context that people use but speech recognizers don’t include:

- The history of the conversation and the topic being discussed.
- Visual cues of the person speaking including facial expressions and lip movement.
- Prior knowledge about the person we are speaking with.

Currently, Android’s speech recognizer has knowledge of your contact list so it can recognize your friends’ names.[4] The voice search in maps products uses geolocation to narrow down the possible points-of-interest you might be asking to navigate to.[5]

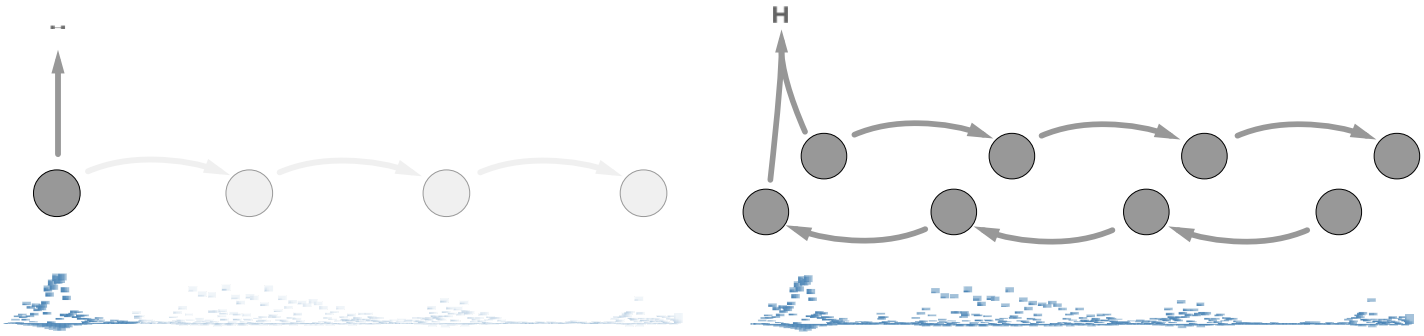
The accuracy of ASR systems definitely improves when incorporating this type of signal. But, we’ve just begun to scratch the surface on the type of context we can include and how it’s used.

Deployment

The recent improvements in conversational speech are not deployable. When thinking about what makes a new speech algorithm deployable, it's helpful to think in terms of latency and compute. The two are related, as algorithms which increase compute tend to increase latency. But for simplicity I'll discuss each separately.

Latency: With latency, I mean the time from when the user is done speaking to when the transcription is complete. Low latency is a common product constraint in ASR. It can significantly impact the user experience. Latency requirements in the tens of milliseconds aren't uncommon for ASR systems. While this may sound extreme, remember that producing the transcript is usually the first step in a series of expensive computations. For example in voice search the actual web-scale search has to be done after the speech recognition.

Bidirectional recurrent layers are a good example of a latency killing improvement. All the recent state-of-the-art results in conversational speech use them. The problem is we can't compute anything after the first bidirectional layer until the user is done speaking. So the latency scales with the length of the utterance.



Left: With a forward only recurrence we can start computing the transcription immediately. **Right:** With a bidirectional recurrence we have to wait until all the speech arrives before beginning to compute the transcription.

A good way to efficiently incorporate future information in speech recognition is still an open problem.

Compute: The amount of computational power needed to transcribe an utterance is an economic constraint. We have to consider the *bang-for-buck* of every accuracy improvement to a speech recognizer. If an improvement doesn't meet an economical threshold, then it can't be deployed.

A classic example of a consistent improvement that never gets deployed is an ensemble. The 1% or 2% error reduction is rarely worth the 2-8x increase in compute. Modern RNN language models are also usually in this category since they are very expensive to use in a beam search; though I expect this will change in the future.

As a caveat, I'm not suggesting research which improves accuracy at great computational cost isn't useful. We've seen the pattern of "first slow but accurate, then fast" work well before. The point is just that until an improvement is sufficiently fast, it's not usable.

The Next Five Years

There are still many open and challenging problems in speech recognition. These include:

- Broadening the capabilities to new domains, accents and far-field, low SNR speech.
- Incorporating more context into the recognition process.
- Diarisation and source-separation.
- Semantic error rates and innovative methods for evaluating recognizers.
- Super low-latency and efficient inference.

I look forward to the next five years of progress on these and other fronts.

Acknowledgements

Thanks to @mrhannun for useful feedback and edits.

Edit

Hacker News discussion.

Footnotes

1. Just ask anyone with a Scottish accent. ↩

2. These results are from Amodei et al, 2016. The accented speech comes from VoxForge. The noise-free and noisy speech comes from the third CHiME challenge. ↩

3. Stolcke and Droppo, 2017 ↩

4. See Aleksic et al., 2015 for an example of how to improve contact name recognition. ↩

5. See Chelba et al., 2015 for an example of how to incorporate speaker location. ↩

Attachment 2

11-3-2017

Automated Speech Recognition for Captioned Telephone Conversations

Jeff Adams CEO
Cobalt Speech and Language, Inc, jeff@cobaltspeech.com

Kenneth Basye PhD
Clark University, kebasye@clarku.edu

Alok Parlikar PhD
Cobalt Speech and Language, alok@cobaltspeech.com

Andrew Fletcher PhD
Cobalt Speech and Language, andrew@cobaltspeech.com

Jangwon Kim PhD
Canary Speech, LLC, jangwon@canaryspeech.com

Automated Speech Recognition for Captioned Telephone Conversations

The State of the Art in 2017 and Projected Paths of Evolution[1]

Authors:
Jeff Adams, CEO, Cobalt Speech & Language, Inc.
Kenneth Basye PhD, Visiting Professor of Computer Science, Clark University[2]
Alok Parlikar PhD, Senior Research Scientist, Cobalt Speech & Language, Inc.
Andrew Fletcher PhD, VP of Research, Cobalt Speech & Language, Inc.
Jangwon Kim PhD, VP of Research, Canary Speech, LLC

Abstract

Internet Protocol Captioned Telephone Service is a service for people with hearing loss, allowing them to communicate effectively by having a human Communications Assistant transcribe the call and equipment that displays the transcription in near real time. The current state of the art for ASR is considered with regard to automating such service. Recent results on standard tests are examined and appropriate metrics for ASR performance in captioning are discussed. Possible paths for developing fully-automated telephone captioning services are examined and the effort involved is evaluated.

1. Introduction

Internet Protocol Captioned Telephone Service (IP CTS) is a service for people with hearing loss, allowing them to communicate effectively by having a human Communications Assistant (CA) transcribe the call and equipment that displays the transcription in near real time.

> *"[IP] CTS allows a person with hearing loss but who can use his or her own voice and has some residual hearing, to speak directly to the called party and then listen, to the extent possible, to the other party and simultaneously read captions of what the other party is saying. In the most common set-up of this service, when an IP CTS user places a call over [an IP CTS] telephone (which is equipped with special software and a screen for displaying captions), the call is automatically connected both to the receiving party (over the PSTN) and via the Internet to a captioned telephone CA."* [FCC17]

In this paper, we describe the current state of the art in ASR as it applies to IP CTS, and discuss the likely paths of evolution for using ASR to assist or replace the human CA in the transcription process.

An ASR system is a complex combination of software and mathematical models of different aspects of speech. In the simplest terms, there are two primary models: the acoustic model, or AM, and the language model, or LM. The AM provides information about how likely it is that a given short segment of audio, say 0.1 seconds' worth, represents speech of a particular kind of

unit. The most often-used units are based on speech phonemes, e.g., the two K consonants and the long vowel A when someone says the word "cake." The LM provides information about how likely it is that a word, or words, will be spoken in conjunction with certain other words. For example, from such a model one might determine how much more likely it is that the word following "ate the chocolate" is "cake" rather than "cape."

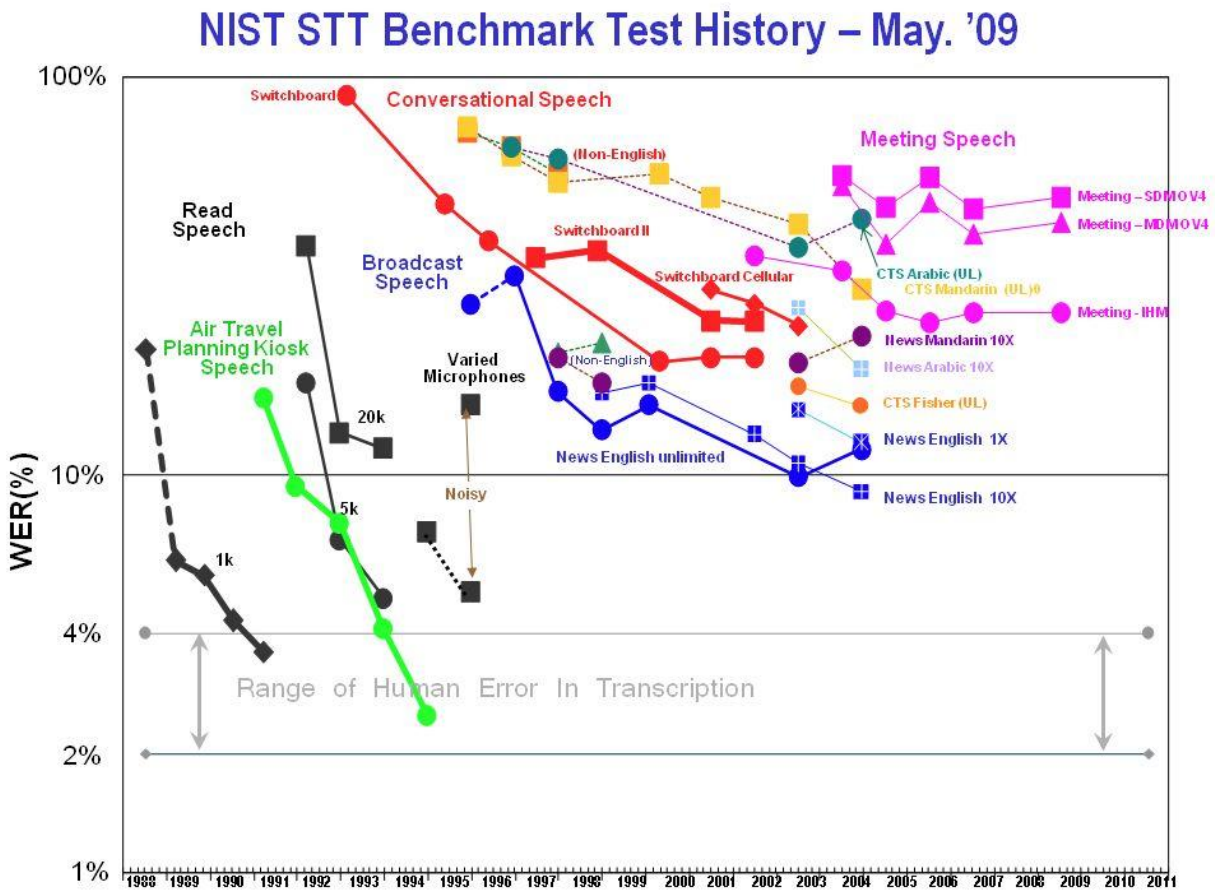
The AM and LM models are connected together by a third element called a "lexicon" that provides the words that may be recognized and modeled by the LM and their pronunciations in the units of the AM. Each of the models is trained on a large collection of data; these collections are called "corpora." For the LM, the training data is in the form of word sequences from real speech or written language; often billions of words or more are used to train an LM. For the AM, the training data is in the form of both speech audio and a detailed transcription; here thousands or tens of thousands of hours of speech are used. Both forms of training data are hard to acquire, making model building a very expensive proposition even to get started. Among several representations commonly used for AMs, deep neural networks, or DNNs, are currently favored since they provide the most accurate models. Training DNNs requires very large amounts of data and is also computationally very expensive. A sophisticated software program called a "decoder" breaks an incoming speech signal in digital form into the short segments used in the AM, and then uses the AM, LM, and lexicon simultaneously to maintain a collection of hypotheses about what the speaker has said, eventually producing a sequence of words which it determines to be the most likely. One very important distinction is between decoders that make this determination at the same rate that people speak versus those that take more time to decode than the time of the spoken audio itself.

2. Theory

In applying ASR to any problem, two fundamental questions emerge. First, what is the nature of the speech provided to the ASR system? Second, what are the relevant measurements of quality and corresponding levels of performance along those dimensions that are required for recognition results? Over decades of research, certain parameters about input speech have been identified as important to performance, including the following: whether the system will be used by only one or by many users, the quality of the audio signal (both the available bandwidth and the presence of noise), whether the speakers have an accent, the age of the speaker, the fluency of the speech, and the context in which the speech is being generated. Among the important contexts that have been considered are speech generated by reading text, speech intended to control an automatic system, speech as dictation, and conversational speech between human speakers.

Depending on the nature of the problem, ASR performance can vary widely. As we will explain below, even the way performance is measured can vary. Most ASR researchers report basic word error rate (WER) figures based on counting all errors in recognition equally. The following figure, put together in 2009 by the National Institute of Standards and Technology (NIST), shows a timeline of ASR error rates for various tasks [PAL03]. It can be observed that (i) for

each task, improvements in speech recognition techniques through years have yielded material performance gains for a given problem; (ii) in the same year, the performance for different problems can be widely different.



The diagram illustrates progress over the years (on the X axis) in terms of WER (on the Y axis, log scale), for a variety of benchmark problems. Note that, though certain benchmarks are used over the course of several years, the data used to measure WER for a particular benchmark was sometimes changed from year to year; this accounts for occasional year-over-year WER *increases* in some benchmarks. On the left, the earliest and simplest benchmarks involved speech that was read by speakers from text, recorded with high-quality microphones in quiet environments, and using a very limited set of possible phrases. One such benchmark envisioned speakers making travel plans using an automated kiosk. The use of speech that was read from text makes the problem artificially easy for ASR because read speech is typically much more carefully pronounced and slower than normal speech. Read speech is also typically better organized, well formed, more predictable, and hence better modeled by the LM. Toward the middle, the blue line represents a benchmark for recognizing spoken news broadcasts. Here the vocabulary and phrasing is near unlimited, the audio is high quality and not noisy, and the speech is generally quite careful, with good articulation by professional announcers. Finally, toward the

right and upper part of the diagram are benchmarks for conversational telephone speech and for transcription of speech from meetings. These represent some of the most challenging speech problems both in terms of having completely open vocabulary and phrasing, highly variable audio quality and noise level, and casual and frequently disfluent pronunciation.

It should be noted that while the chart above identifies the "range of human error in transcription" as 2-4%, that was not based on realistic measurements, and more recent estimates of human transcription error have put the number at 5% and higher. The actual number varies depending on the skill of the transcriber, the clarity of the audio, and the accuracy required by a given task.

This image is the latest such benchmark that NIST has made available, and although it is now a bit dated, it throws light on how ASR might be a "solved" problem for one domain, but not for others. In our experience, the situation shown above still holds today. That is, many standard research benchmarks are not solved in the sense of having reached parity with humans.

Recent years have seen the rise of natural-language conversational assistants, such as Apple's Siri, Amazon's Alexa, Microsoft's Cortana, and Google's "OK Google" assistant. These constitute an ASR task not represented by the various benchmark problems tracked by NIST. They carry their own set of challenges, but are also less demanding in terms of accuracy, since the intent of a command may be correctly recognized, even if some of the words of the command are misrecognized by the ASR.

3. Results

In terms of applying ASR to IP CTS, full automation would replace the human CA with an ASR system that could automatically generate transcripts of equivalent (or better) quality. Considering the question of the nature of the speech provided, we would expect conversational speech, from a variety of speakers, over fixed and mobile telephone service, with a variety of background noise conditions, accents, ages, and fluency levels. The speech requiring accurate, real-time transcription could vary from quite formal to very casual, and there are no limits on what might be said. This presents us with one of the most difficult areas of the ASR problem space.

One way to think about how ASR can be applied to IP CTS is to note that caption quality is ultimately determined by the usefulness of the service to the end-user in facilitating communication. Further, the key parameters for quality fall into three categories: accuracy, latency, and readability. Obviously, for any communication, accuracy is of paramount importance, although it is important to acknowledge that not all recognition errors have the same impact. For example, a misrecognition of "She hasn't been here" resulting in "She has been here" will almost certainly be much worse than a misrecognition resulting in "She hasn't been hear." Current standards for IP CTS divide errors into "major" and "minor" with only major errors considered in quality metrics. Unfortunately, this kind of measurement is particularly difficult to automate because determining whether an error is important to understanding can

currently only be done reliably by humans. As we mentioned above, most ASR researchers report basic word error rate (WER) figures based on counting all errors without regard to their potential impact on understanding, so traditional WER numbers can be at best only an approximate guide for IP CTS accuracy.

The latency of an IP CTS system is also very important. Real-time conversations proceed smoothly when latencies are very low but become more difficult as they increase, so that the benefit of IP CTS to users with some hearing decreases if the transcription has a long lag relative to the audio. IP CTS latencies are comparatively easy to measure, and are currently in the range of a few seconds, though reducing this lag is a stated – and appropriate – goal of the FCC [FCC16].

Readability impacts the usefulness of transcription in real-time conversation. Studies show that having transcriptions that include correct punctuation and capitalization reduce the cognitive load required to read them [JON03]. One particularly important form of punctuation is the indication of a question. In vocal communication, this is often suggested only by a change in intonation, as in "You know what to do?" In the context of IP CTS, speakers may be relying entirely on these changes in intonation to convey their meaning, and the presence or absence of a question mark may represent a significant difference in meaning, not simply an improvement in readability. Measuring readability is currently time-intensive since it usually involves human subjects, but there are good proxies that could function automatically.

3.1 Applying ASR to IP CTS in practice

Considering the three quality parameters important for IP CTS described above, the current state of the art in ASR achieves various levels of performance against each parameter. Markedly wide variability of conditions exist which ASR systems must deal with and as a result, no one system performs well against all three.

For example, current dictation systems like Nuance's Dragon NaturallySpeaking generate very readable transcriptions. Users can either vocalize punctuation and capitalization cues or the system can infer and insert them automatically. Although automatic punctuation insertion can be a very useful feature, users typically have a mixed experience; it appears that even with auto-punctuation turned on, users have to adapt their dictation style to suit the type of speech that the auto-punctuator expects, and most users tend to vocalize punctuation, which is not natural to do in the context of a telephone conversation. Latency is very low; words appear as they are spoken with minimal delay. Accuracies can be quite high for high-quality audio and particularly for speakers with considerable experience; this is both a matter of adaptation of speaker-specific models on the part of the ASR system and of learning on the part of the speaker, who gets constant feedback in the form of results and unconsciously learns how to speak so as to be correctly understood. This is why the Nuance product is utilized by some IP CTS provider CAs in the delivery of captions in IP CTS. After the CA trains a profile to their voice, during a live call, they revoice what the IP CTS caller says. The CA monitors the text output delivered to the

IP CTS user and makes manual corrections as needed. The use of Nuance is faster and more accurate than traditional typing done by a CA. However, using a dictation system like this directly on casual conversational speech consisting of multiple unknown users talking on the telephone, rather than as a tool used by the CA to create the captions, will have low accuracy and is infeasible at present for the purpose of creating an accurate, readable transcript.

Similarly, recently introduced conversational assistant systems like Google Now, Apple's Siri, and Amazon's Echo/Alexa deal well with noisy environments and fairly casual speech. Latencies are also low here. But since these systems are built for command and control, readability of transcriptions isn't a design goal and the range of things that can be correctly recognized is generally limited to specific command grammars [JAC17, MAR17]. The problem addressed by these systems is also very different in terms of speakers. A "personal assistant" (PA) application can (and certainly does) take advantage of the fact that it is generally used by one person, or at most a few people, almost exclusively. This means that adapting the recognition models for that speaker or speakers can improve overall performance very considerably. Additionally, since speakers are highly motivated to make themselves understood correctly, they can again be expected to self-train as they use the system. Finally, these systems can still be a commercial success even if they don't work that well for a significant fraction of the population as long as they work well for most people. The contrast between the PA application and IP CTS is sharp. PA systems aren't designed to recognize and transcribe conversational speech between two humans. They rely heavily on being able to adapt to one or a small number of speakers who are speaking intentionally to be understood by the system. IP CTS systems, on the other hand, must deal well with many disparate speakers, some only very infrequently, who are speaking to another human. The conclusion is that recent progress in accuracy of PA systems does not translate well into accuracy for IP CTS systems.

Within the ASR research community, work and systems focused specifically on conversational telephone speech have been around for about two decades. One of the first and most widely used Speech corpora for this area is the Switchboard corpus [GOD92]. Naturally, performance on this corpus is interesting when evaluating the feasibility of using ASR for IP CTS.

"Switchboard" is a collection of about 2,400 two-sided telephone conversations among 543 speakers (302 male, 241 female) from all areas of the United States. A computer-driven robot operator system handled the calls, giving the caller appropriate recorded prompts, selecting and calling another person to take part in a conversation, introducing a topic for discussion and recording the speech from the two subjects into separate channels until the conversation was finished. About 70 topics were provided, of which about 50 were used frequently. Selection of topics and callees was constrained so that: (i) no two speakers would converse together more than once and (ii) no one spoke more than once on a given topic. Switchboard is a useful tool, but not fully representative of conversational speech on the telephone since there is no overlapping of voices, speakers were not known to each other, and only a finite number of topics are represented.

Last year, Microsoft published promising results on the Switchboard data: their recognition system produced output that had a WER of 5.9% [XIO17]. They claimed that at this error threshold, their system was “just as good as humans are in recognizing speech”. Recently, IBM has also published results that show a WER of 5.5%. However, IBM has also acknowledged that this result does not imply speech recognition has achieved parity with human performance. They claimed that although some believe that human parity is achieved at an error of 5.9%, their newest experiments show that parity is achieved at a much lower error of 5.1%. Thus, even their "best" system would still need a 0.8% reduction in order to achieve human-like performance in a WER sense. That may not seem like a large difference, but it represents a 14% relative reduction in the number of errors, which is a significant gap.

The absolute numbers – whether human parity can be achieved at 5.9% or 5.1%, or some other threshold – is one part of the story. There is another facet of practicality of these systems. These "best" ASR systems are built as large recognition models, by throwing large amounts of data and computation resources at the problem. In fact, the ASR system isn't always one engine, but a combination of multiple engines that run recognition on the same audio in parallel, and another engine that combines these different outputs to produce something that is better than any individual system. This represents a nice technical feat, but by its very nature implies a system that can run recognition multiple times with different engines and/or models, and then pick and choose the best results for a final decision, which requires time to process after the speech ends. The time needed for these systems to process the speech to get the advertised accuracy was not made public, but it is the estimate of the authors of this paper that it is measured in minutes or hours, not milliseconds, and is impractical for an IP CTS system.

Such systems, with their multi-pass approach and dependence on very large computing platforms, prioritize accuracy at the cost of latency and speed. They are great experimentation platforms but are not commercially viable for problems with real-time requirements such as IP CTS. It is likely, therefore, that significant forward progress in WER for ordinary telephone conversations is likely to be asymptotic in nature, requiring technical breakthroughs and exponential compute power increases to achieve a comparatively few percentage points of required accuracy.

There is also a case to be made about the evaluation methodology itself. Most of the "top" results on the Switchboard data have been reported on a newer test set, and the documentation of that set mentions that there is some overlap with the speakers found in the training data [NIST00]. In real-life situations, most of the speech recognition run in the context of IP CTS would be on speakers not seen in the training data. If all speakers in the Switchboard testing data were "unseen" from the data that these models are trained on, there is reason to believe that the performance of these top systems, in absolute terms, would be one of a higher error rate.

Claims of achieving parity with human performance or 100% accuracy make good press for

research organizations. But such claims, even if they were indisputable (and they certainly are not) aren't a reliable basis on which to judge the possibility of a fully automated IP CTS system. It is important to keep in mind that transcription accuracy in IP CTS and transcription accuracy as measured by ASR researchers are not the same thing. On one hand, IP CTS accuracy measures don't count minor errors, which would seem to make the problem simpler. On the other hand, IP CTS transcription is done live, based on only one opportunity to listen to the data, with a requirement that the transcript be generated as immediately as possible. The most-often used accuracy level for current IP CTS systems is that transcriptions should be 98% free of major errors. Even high-latency systems like those discussed above don't come close in terms of simple WER; to our knowledge no one has evaluated their Switchboard results in terms of the major/minor error distinction used for IP CTS.

4. Discussion

For IP CTS applications, there are four broad challenges that will need to be surmounted for widespread commercialization. First, and most fundamental, is that accuracy on casual conversational telephone speech will need to be improved considerably from currently achievable levels. Second, the readability of transcripts will need to be improved from what can currently be done automatically for conversational speech. Specifically, punctuation and capitalization are required to make transcripts easily readable. Third, the latency of transcription must be kept low and, if possible, reduced even from current levels. Finally, all the computation required must be done without requiring excessive computational resources.

One unusual aspect of the IP CTS problem is that the "user" of the system is the person reading the transcripts, but the "speaker" is someone else and there are, in fact, many speakers even for a single user. Also, because of factors like background noise, separate calls from the same speaker may be quite different. In a usual context, one ordinarily asks how much effort would be required to make the system work well for, say, 95% of users in 90% of their uses. But for IP CTS, the right question to ask is how much effort is required to make the system work well for some percentage of calls. Rather than waiting until the system reaches some threshold percentage and deploying it for all calls, it could make sense to build a hybrid system using both ASR and human transcribers, wherein the system identifies calls as suitable for ASR transcription, with human CAs handling the remainder. Some research would be required to determine a suitable method for identifying a call (or a portion of a call) as suitable for ASR. This is a challenging problem that has not been addressed by prior published research. The viability of such a "selective" approach to ASR for IP CTS depends therefore on both improvements in ASR accuracy, as well as improvements in the ability to model which calls are suitable for ASR.

We expect that the pathway to a commercialized, fully automated IP CTS system involves many iterations of a cycle well known to ASR system-builders. The cycle starts with the acquisition of massive quantities of speech data, preferably from a source matching the characteristics of the speech and language of the target application and the associated transcriptions. The data are

used to build models, the models are evaluated, and the system is improved by the addition of more data or by employing new modeling and/or decoding techniques, and this process is repeated until the system achieves an acceptable standard of quality. The availability and compilation of such data sets required to train deep neural networks is itself a key barrier. Capturing vast quantities of conversations is either labor intensive, or requires overcoming privacy concerns among a large population of contributors – for example, providers are restricted in recording IP CTS calls, making it difficult to create a realistic data set. Innovations will be required in methods for training speech models from encrypted and/or anonymized data in order to make substantial improvements in ASR accuracy for IP CTS.

Keeping in mind the need for real-time transcription, our consensus estimate is that a system based on the current (2017) state of the art for conversational telephone speech would work for fewer than 10% of all calls. A concerted effort started now by a team of ten experienced speech researchers and engineers might improve this by 6-8% (absolute, not relative) for four to six years, reaching a system which could handle roughly 50% of all calls. The need to be conservative in identifying which calls can be safely handed off will reduce this number somewhat, and system elements will need to be developed and tested that enable switching to human assistance when machine performance falls short, perhaps at the request of the IP CTS user. Future years would likely see continued progress but at a somewhat slower rate, perhaps reaching the ability of handle 75% of all calls in another four to six years. The most difficult calls will take even longer to handle reliably. Waiting to start the effort would shorten these times somewhat as ASR research can be expected to improve the performance at the "starting point," but the IP CTS problem is both sufficiently hard and has enough unusual requirements that the effort will remain considerable for the foreseeable future.

Finally, it should be noted that commercial viability of any advances in ASR are affected by market and financial incentives, or the lack thereof. In the case of PA systems the financial incentives are clear and immediate while financial incentives to drive progress in ASR systems addressing conversational speech, such as IP CTS, are less obvious. In the NIST graph presented earlier, one might note that progress was made more rapidly on some projects than on others. This effect is largely due to the financial incentives applied to a particular ASR task. Substantial government funding was provided to groups working on tasks like Switchboard and Broadcast News, and commercial incentives have spurred remarkable progress in conversational assistants such as Siri and Alexa. At the same time, there has been little or no funding for improvements in meeting transcription, leading to the relatively stagnant progress shown in the NIST graph. From this, we can infer that future progress in ASR for IP CTS will be dependent on the availability of funding from either government or commercial sources.

5. Conclusion

Automatic Speech Recognition has made considerable progress on the very difficult problem of recognizing human speech. Some systems achieve human-level performance on fairly narrow

tasks and recent advances by research groups have done fairly well on quite difficult tasks like recognizing conversational speech, though there are many reasons to doubt claims of reaching human-level performance. Internet Protocol Captioned Telephone Service represents one of the hardest problems for ASR. This is true both in the sense of the speech involved, which is conversational, can be quite noisy, and presents many different speakers for short durations, and in sense of the performance requirements, which include very high accuracy, very low latency, and an additional requirement of generating easily read transcriptions. Although it's possible, and perhaps even likely, that ASR will improve to the point that a fully automated IP CTS system can be made commercially viable, our belief is that that point is still well into the future.

References

[FCC17] "Internet Protocol (IP) Captioned Telephone Service" FCC Consumer Guide, 2017 https://www.fcc.gov/consumers/guides/internet-protocol-ip-captioned-telephone-service

[FCC16] "Transition From TTY to Real-Time Text Technology" Federal Register, 81 FR 33170, 2016 https://www.federalregister.gov/documents/2016/05/25/2016-12057/transition-from-tty-to-real-time-text-technology

[GOD92] Godfrey, John J., Edward C. Holliman, and Jane McDaniel. "SWITCHBOARD: Telephone speech corpus for research and development." Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on. Vol. 1. IEEE, 1992.

[JAC17] Purewal, Sarah Jacobsson, and Cipriani, Jason. "The Complete List of Siri Commands" CNET Mobile blog, 2017. https://www.cnet.com/how-to/the-complete-list-of-siri-commands/

[JON03] Jones, Douglas A., et al. "Measuring the readability of automatic speech-to-text transcripts." INTERSPEECH. 2003.

[MAR17] Martin, Taylor, and Priest, David. "The Complete List of Alexa Commands So Far." CNET Smart Home blog, 2017. https://www.cnet.com/how-to/amazon-echo-the-complete-list-of-alexa-commands/

[NIST00] "The 2000 NIST Evaluation Plan for Recognition of Conversational Speech over the Telephone" http://www.itl.nist.gov/iad/mig/tests/ctr/2000/h5_2000_v1.3.html

[PAL03] Pallett, David S. "A Look at NIST's Benchmark ASR Tests: Past, Present, and Future." National Institute of Standards and Technology (NIST), 2003 http://itl.nist.gov/iad/mig/publications/ASRhistory/index.html

[SAO17] Saon, George et al. "English Conversational Telephone Speech Recognition by Humans and Machines." 2017 arXiv:1703.02136 [cs.CL]

[XIO17] Xiong, Wayne et al. "Achieving Human Parity in Conversational Speech Recognition" 2016 arXiv:1610.05256 [cs.CL]

[ZEK17] Zekveld, Adriana A. et al. "User Evaluation of a Communication System That Automatically Generates Captions to Improve Telephone Communication." Trends in Amplification 13.1 (2009): 44–68. PMC. Web. 4 May 2017.

Endnotes

1. While this paper was sponsored by CaptionCall, Inc., the thoughts and opinions expressed here represent the independent views of the authors.

2. Corresponding author. kebasye@clarku.edu Department of Mathematics and Computer Science, Clark University, 950 Main St. Worcester, MA 01610.